# 基于梯度提升决策模型的空间占用检测研究 *

徐新卫[1,2]，丁敬安[1]，柳智才[1]，王多梅[3]，腾　翔[1]，邵瑞瑞[1]

(1. 安徽工业大学 管理科学与工程学院，安徽 马鞍山 243000; 2. 南京大学 计算软件新技术国家重点实验室，南京 210000; 3. 河海大学 公共管理学院，南京 210000)

**摘　要：**随着绿色建筑和绿色生态城区经济激励机制基本形成，面对大量多维空间占用数据，“大数据绿色建筑”节能体系应运而生。然而大量多维的建筑数据却没有被充分利用，且传统空间占用检测模型分类精度还不够准确，模型时间复杂度较高。利用 UCI 占用检测数据集，在原始数据集上加入时间戳，使模型分类精度均获得提高，同时利用 MCMR（最大相关最小冗余）方法进行特征选择，通过随机森林作为分类器验证分类效果，获取最优特征子集。且利用选取的特征子集构建占用检测模型，其中 XGBoost 模型与随机森林模型（RF）进行比对，分类精度较高，且时间复杂度更低。

**关键词：**大数据绿色建筑；空间占用检测；最大相关最小冗余；梯度提升算法

**中图分类号：**TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.09.0907

## Occupancy detection based on extreme gradient boosting decision model

Xu Xinwei[1, 2], Ding Jingan[1], Liu Zhicai[1], Wang Duomei[3], Teng Xiang[1], Shao Ruirui[1]

(1. *School of Management Science & Engineering Anhui University of Technology*, *Maanshan Anhui* 243000, *China*; 2. *State Key Laboratory for Novel Software Technology*, *Nanjing University*, *NanJing* 21000, *China*; 3. *School of public administration*, *Hohai University*, *NanJing* 210000, *China*)

**Abstract:** With the green buildings and green-economic environmental cities are gradually formed, "big data green building" energy conservation systems come into being. However, a large number of multi-dimensional building data are not fully utilized and occupancy detection with accuracy of traditional algorithms is not accurate with the higher time complexity. This article acquired the data of Occupancy Detection from UCI. Add a timestamp to the original dataset, the accuracy is increased. Using the MCMR method to select features with maximum correlation and minimum redundancy, random forest is using as classifier to verify classification effect . The XGBoost model constructed by the optimal subset is compared with the random forest model (RF) , and the classification accuracy is higher and the time complexity is lower.

**Key Words:** big data of green buildings; occupancy detection; MCMR; XGBoost



2017 年 3 月 1 号，《住房城乡建设事业“十三五”规划纲要》提出“发展绿色建筑、绿色建材，大力强化建筑节能”，明确要求“到 2020 年，城镇新建建筑中绿色建筑推广比例超过 50%，绿色建材应用比例超过 40%，新建建筑执行标准能效要求比‘十二五’期末提高 20%”。3 月 21 日，以“提升绿色建筑质量，促进节能减排低碳发展”为主题的第十三届国际绿色建筑与建筑节能大会暨新技术与产品博览会在北京国家会议中心召开，会议上除了从国家政策、经济形势、产品提升、运营管理等方面探讨，技术创新、平台建设与大数据分析等方面成为建设绿色建筑与建筑节能的热点话题。由于我国人口众多，资源、能源有限，能源、经济和环境问题已经成为一个城市的热点话题,如何合理的综合利用能源,提高城市能源利用效率、优化资源配置、保护人类赖以生存的自然环境已然是当今社会关注的热点问题。同时随着“大数据”时代的到来，绿色建筑发展理念[1,2]和数据挖掘、机器学习等技术的结合得到了一定的发展，通过对大量多维异构建筑空间占用数据特征提取，寻找最优特征子集构建模型，进行空间占用检测，提高检测的分类精度，是对建筑内部相关能源优化配置的一种新方法。

空间占用检测本质属于模式识别范畴[3~5]，利用多传感器监测室内环境获取空间占用，通过特征的提取和分类算法的选择

基金项目：国家社科基金资助项目（15BJL014）

作者简介：徐新卫（1971-），男，湖北武人，副教授，博士研究生，主要研究方向为智能计算、数据挖掘；丁敬安+（1991-），男（通信作者），安徽宿州人，硕士研究生，主要研究方向为数据挖掘、人工智能及模式识别（986462579@qq.com）；柳智才（1993-），男，安徽阜阳人，硕士研究生，主要研究方向为人工智能及模式识别；王多梅（1995-），女，安徽阜阳人，硕士研究生，主要研究方向为资源与环境管理；腾翔（1992-），男，江苏盐城人，硕士研究生，主要研究方向为光伏发电与并网计算；邵瑞瑞（1991-），女，安徽淮南人，硕士研究生，主要研究方向为复杂网络在社交网络中的应用研究.

进行分类模型的构建，对空间占用状态进行预测，利用预测的准确性进行 *HVAC* 系统智能控制，达到节省能源耗费的目的。研究表明通过空间占用检测技术对 *HAVC* 进行智能控制，理论上估计能够年节省能源 29%~80%左右[6~10]。通过摄像机与传感器引入时间特征，获得实时数据，作为扩展卡尔曼滤波算法模型输入值，能够提高空间占用检测准确率[11]。利用 *RFID* 技术和运动传感器进行实时空间占用检测能够减少天然气消耗和空间被占用但房间不温暖的时间，从而提高空间的舒适度[12]。通过不同的特征组合与算法模型的结合，对空间占用检测的准确率也有一定的影响。提取商业建筑现有的环境资源，包括进入许可证、无线记录，日程表，通信客户端等获得数据，使用线性回归和 C4.5 算法进行占用检测准确率达到 90%左右[13]。通过决策树对单传感器数据进行占用检测得到的准确率是 97.9%，进行多传感器数据特征组合检测的准确率可达 98.4%。而加入声音和 $CO_2$ 等传感器数据时，预测的结果却不太理想，得出决策树可以提高单一传感器检测准确率，对过多传感器数据特征组合检测会出现精度下降现象[14]。通过对光线、温度、湿度、$CO_2$ 等传感器数据特征组合进行空间占用检测时，发现随机森林在对全部特征组合进行预测时，出现过拟合现象。而线性判别分析在仅只有两个特征的准确率可以达到 97%，不同的特征组合对预测的精确度有影响[15]。通过上述文献分析，其一，可以发现引入时间戳能够提高空间占用检测的准确率；其二，多传感器特征组合，却导致分类模型分类精度下降，分析主原因是，特征与特征之间具有高度相关性，或者存在冗余特征。某些特征包含类别信息量较少，对分类识别效果很低，影响模型的分类性能和时间复杂度。

针对多传感器特征组合出现模型分类精度下降以及时间戳引入提高检测率等问题，本文提出了一种最大相关最小冗余 *MCMR*（maximum correlation and minimum redundancy）的特征选择算法。在 *UCI* 占用检测原始数据集上提取时间变量，细化时间粒度从而引入时间戳形成新的数据集，在新的数据集上利用 MCMR 进行特征的选择，选择出的最优特征子集作为梯度提升决策模型（*XGBoost*）算法的输入。*XGBoost* 是一种梯度提升框架下的集成学习算法，具有灵活可移植的分布式决策梯度提升库，在处理大量数据时，保证相对较高分类精度下，时间复杂度更低[16-18]。在分类研究中，采用该方法对空间占用信息提取进行详细研究成果较少。本文结合了 *MCMR* 进行特征选择的优点与 *XGboost* 分布式并行运算的优点相结合，从特征组合的角度达到了提升空间占用自动识别模型的准确率，同时降低了模型的时间复杂度。

本文主要贡献如下，在原始数据的基础上，加入时间戳，改变 *XGBoost* 和 *RF* 算法无法处理时间变量。实验结果表明，加入时间戳与没有加入时间戳的模型相比分类精度均得到提高，变化最明显的是 *XGBoost* 在 *testing* 测试集上分类精度提高了 4.09%，*RF* 在 *testing* 数据集上分类精度提高了 2.78%。同时，*RF_1* 比文献 19 引入时间戳更合理。

利用 *MCMR* 特征选择方法进行特征选择，剔除关联度小冗余度高的 *HumidityRatio* 特征，利用随机森林作为分类器，进行迭代寻优，获取了最优特征子集。通过特征与分类算法构建检测模型得到 *XGBoost* 在训练样本数据集上的分类精度最高，精度为 99.41%；*SVM* 在测试样本数据集 1 上的分类精度最高，精度为 97.90%；BP 在测试样本数据集 2 上的分类精度最高，精度为 99.07%。最后将 *XGBoost* 与随机森林（*RF*）分类方法进行比较，*XGBoost* 模型分类精度更高，算法时间复杂度更低。评估基于 *XGBoost* 算法的多传感器数据源综合分类方案在空间占用分类中更实用，为 HAVC 系统智能控制及构建绿色建筑节能体系提供依据。

## 1　理论与方法

### 1.1　MCMR 特征选择方法

基于关联性较小冗余度较高的特征，影响分类模型的精度，所以需要对样本数据集进行特征筛选。特征选择目的在于从样本数据集合中选择一个规模较小的特征子集，该子集能够在数据挖掘和机器学习任务中提供与原集合近似或者更好的表现。在不改变特征包含类别信息量的基础上，较少特征为数据提供了更强的可解读性[19-20]。传统特征选择方法中，特征之间的相关性只考虑特征之间的线性相关或非线性，没有考虑特征之间的全相关性，再者特征选择往往将特征相关性和冗余性分割判断，无法判断整个特征子集的组合效应[21-22]。

以线性相关和非线性相关为基础，计算特征之间的全相关系数度量特征间的独立和冗余程度。同时以信息论为基础，计算特征与类别间的互信息，即特征含有类别信息量的大小，表示特征与类别间的关联程度[23-27]。综合封装式和滤波式两种特征选择方法的优点，本文提出一种最大相关最小冗余 *MCMR*（Maximum Correlation and Minimum Redundancy）的特征选择算法。

假设样本集中想 $x$，$y$ 分别表示长度为 $n$ 的成对连续变量，通过 *pearson* 相关，计算特征之间的线性相关系数 r，如式（1）所示。

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2} \tag{1}$$

构造关联度矩阵 $A$。利用距离相关，计算特征之间的非线性相关系数 $R_n^*$，构造关联矩阵 $B$，$dcov(x,y)$为变量 $x$，$y$ 的距离协方差，$dvar(x)$与 $dvar(y)$分别为变量 $x$，$y$ 的距离标准差，如式（2）所示。

$$R_n^* = dcor(x,y) = \frac{dcov(x,y)}{dvar(x)dvar(y)} \tag{2}$$

综合线性相关和非线性相关，计算特征之间的全相关，得到关联矩阵 $C$。特征之间的全相关计算过程如下，全相关系数为 $w$。

$$w = e^{[e^{|r|-1}+e^{R_n^*-1}]-2} \tag{3}$$

其中：$i$，$j$ 代表关联矩阵中第 $i$ 为 $j$ 列，得到全相关矩阵 $C$。

以信息论为基础，计算特征与类别间的互信息用 $P(x_i)$ 表示特征 $x$ 取第 $i$ 个值 $x_i$ 的概率，$P(x_i|y_j)$ 表示类别 $y$ 取值为 $y_j$ 时特征 $x$ 取值为 $x_i$ 的概率。$x$ 的信息熵 $H(x)$ 及已知变量 $y$ 后 $x$ 的条件信息熵 $H(x|y)$ 的计算方法如下：

$$H(x) = -\sum_i p(x_i) logp(x_i) \tag{4}$$

$$H(x|y) = -\sum_j p(y_j) \sum_i p(x_i|y_j) logp(x_i|y_j) \tag{5}$$

变量 $x$、$y$ 之间的互信息 $MI(x，y)$ 可按以下公式计算：

$$\begin{aligned} MI(x,y) &= H(x) - H(x|y) = H(y) - H(y|x) \\ &= \sum_{x,y} p(xy) \log \frac{p(xy)}{p(x)p(y)} \end{aligned} \tag{6}$$

然而互信息存在偏好等问题，所以本文采用互信息率来度量特征 $x$ 与类别 $y$ 间的相关性：

$$sim(x, y) = \frac{MI(x, y)}{H(x)} \tag{7}$$

由此得到的相关度 $sim(x，y)$ 的取值范围在[0,1]，相关度为 0 表示两个特征不相关，相关度为 1，则表示两个特征完全相关。

MCMR 特征选择算法描述如下：

输入：训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$

a) 离散化及初始：采用 ChiMerge 方法离散数据集 D 中的连续特征，结果仍用 D 表示；

b) 根据式（1）～（2）计算数据集 D 中任意两个特征之间的全相关系数 $w$，根据式（4）～（7）计算数据集 D 特征类别之间相关系数 $sim$；

c) 设置参数 $\alpha \in [0,1]$，找出 $w > \alpha$ 相关特征，比较 $sim$ 值；

d) 删除 $sim$ 值较小的特征，通过随机森林分类器精度验证合理性；否则，*updata* $\alpha = \alpha + 1$，回到第 3 步；

e) end for.

输出：特征子集 $d = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

### 1.2　梯度提升算法分类机制

*XGBoost*（extreme gradient boosting，*XGBoost*）是一种基于 *GBDT*（gradient boosting decision tree，*GBDT*）梯度下降框架的集成学习算法。GBDT 是“梯度下降”和决策树相结合，基于前一个模型残差减少的方向上，构造新的分类器，依次迭代，构造一组弱分类器，弱分类器输出结果进行加权累加作为强分类器输出结果[28-29]。*XGBoost* 与 *GBDT* 区别在于，改变了 *GBDT* 基于 *Boosting* 串行序列化求解问题，利用 *CPU* 多线程分布式并行计算，并通过对残差进行泰勒二次展开进行求解，从而打破现有库的计算速度和精度，使得数据处理和运算的速度得到了提升。

训练 *GBDT* 分类算法基本步骤如下：

输入：训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$

a) 初始 $F^*(x) = argminE_y[L(y, F(x))|x]$；

b) $F_0(x) = f_0(x)$；

c) for m=1,2,…,M do；

(a) $g_m(x) = -\frac{\partial E_y[L(y, F(x))|x]}{\partial F(x)}|_{F(x)=F_{m-1}(x)}$；

(b) $\rho_m = \arg\min E_y[L(y, F_{m-1}(x)) + \rho g_m(x) | x]$；

d) Update $F_m(x) = F_{m-1}(x) + \rho_m g_m(x)$；

e) end for.

输出　$F^*(x) \approx F_m(x) = f_0(x) + \sum_{m=1}^{M} \rho_m g_m(x)$

上述步骤中，$F^*(x)$ 为寻找使得期望损失最小的决策函数，$L(y,F(x))$ 为损失函数，$g_m(x)$ 为当前模型的负梯度方向，$\rho_m$ 为计算损失函数的负梯度在当前模型的值，将它作为残差的估计，估计回归树叶节点区域，以拟合残差的近似值，利用线性搜索估计叶节点区域的值，使损失函数极小化，更新回归树，得到输出的最终模型 $F_m(x)$。

*XGBoost* 对损失函数 $obj^{(t)}$ 做了二阶的泰勒展开，并在目标函数之外加入正则项 $\Omega(f_t)$，整体求最优解，用以权衡目标函数的下降和模型复杂程度，避免过拟合。式（8）中将目标函数做泰勒展开，并引入正则项：

$$\begin{aligned} obj^{(t)} &= \sum_{i=1}^{n} l(y_i, y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant \\ &\approx \sum_{i=1}^{n} [l(y_i, y_i^{(t-1)}) + \partial_{y_i^{(t-1)}} l(y_i, y_i^{(t-1)}) f_t(x_i) + \\ &\quad \frac{1}{2} \partial^2_{y_i^{(t-1)}} l(y_i, y_i^{(t-1)}) f_t^2(x_i)] + \Omega(f_t) + constant \end{aligned} \tag{8}$$

求解每个样本的一阶导 $g_i$ 和二阶导 $h_i$，将目标函数按叶子节点归约分组得如下公式：

$$\begin{aligned} obj^{(t)} &\approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \\ &= \sum_{j=1}^{T} [G_i w_i + \frac{1}{2}(H_i + \lambda w_i^2)] + rT \end{aligned} \tag{9}$$

## 2　空间占用检测模型构建流程

本文对空间占用样本数据集加入时间戳，利用 *MCMR* 特征选择法，删除关联度低，冗余度高的特征，选取最优特征子集，利用最优特征子集以及 *XGBoost* 分类算法构建空间占用检测模型，流程图如图 1 所示，主要包括以下几个步骤：

a)加入时间戳。通过对原始数据集日期变量进行重新提取，增加分类特征，改变原有算法模型不能直接处理时间变量，对空间占用进行实时检测，构建样本数据集。

b)利用 *MCMR* 方法选择特征。通过计算特征之间全相关系数，及特征与类别之间的互信息率，选择相关度较高冗余度较低的特征。

c)特征子集选取及验证。利用封装式随机森林特征递归删减法，验证 *MCMR* 特征选择方法的合理性，获取最优特征子集。

d)训练分类器构建。输入上一步选取的特征子集作为训练样本，通过迭代建立一系列回归决策树，构成 *XGboost* 分类器

学习模型。

*e)HAVC* 系统智能控制。通过分类器模型的学习，获得较高分类精度的模型。依据学习逻辑对现有室内环境变量因素分析，预测占用状态，智能调节 *HAVC* 系统，达到节省能耗的目的。

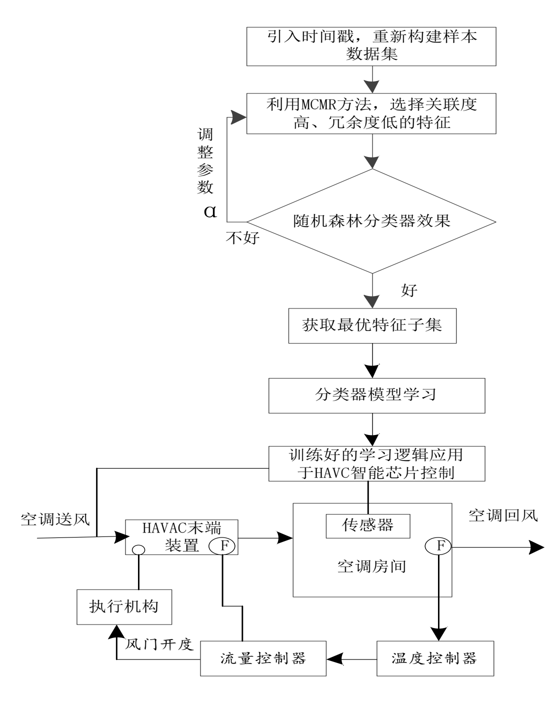



图 1 空间占用检测模型构建流程图

## 3　实验设计与结果分析

### 3.1　实验数据

空间占用检测影响因素往往存在检测困难等问题，所以只能获得较容易监测数据，导致大体量超高维数据很难获得，本文的数据来源于 *UCI* 上 *Occupancy Detection* 数据集，其中训练集 training 和测试集 *testing* 均是门关闭时测量得到，*testing2* 测试集是门打开时测量得到，数据集包含的变量为：时间(*Date*)、温度（*T*）、湿度（*H*）、光照（*Light*）、$CO_2$ 浓度（$CO_2$）、湿度比（*HR*）和占用状态（*Occupancy*），*training* 为 8143 条记录，*testing* 为 2665 条记录，*testing2* 为 9752 条记录。其中 *training* 数据格式如表 1 所示。

表 1　实验数据集

| Date | T | H | Light | CO2 | HR | Occupancy |
|---|---|---|---|---|---|---|
| 2015/2/4 18:04 | 23 | 27.125 | 419 | 686 | 0.004714942 | 1 |
| 2015/2/4 18:06 | 23 | 27.125 | 418.5 | 680.5 | 0.004714942 | 1 |
| 2015/2/4 18:07 | 23 | 27.2 | 0 | 681.5 | 0.004728078 | 0 |
| 2015/2/4 18:08 | 22.945 | 27.29 | 0 | 685 | 0.004727951 | 0 |
| 2015/2/4 18:08 | 22.945 | 27.39 | 0 | 685 | 0.004745408 | 0 |

### 3.2　评价指标

对于空间占用检测模型，本文选用决策模型常用的混淆矩阵，作为模型性能评价指标，对训练样本集和测试样本集进行性能度量。

使用 2×2 的混淆矩阵来表示预测准确度：

$$\text{准确度} = \frac{TP + TN}{TP + TN + FP + FN}$$

其中：*TP*、*TN*、*FP* 和 *FN* 指的是模型预测值落入这些类别中的次数，因此，准确度表示正确分类的数目除于所有预测值的个数。

### 3.3　实验与分析

本文实验环境如下：操作系统为 Windows 7，CPU 为 Intel® Core™ i5-3210M @2.5GHz，实验内存为 4GB，主要实验平台 R version 3.3.3 版本。

1）数据预处理

（1）引入时间戳

由于梯度提升模型和随机森林模型无法对时间变量直接处理，所以本文对收集到的数据进行如下处理，利用 *lubridate* 包对时间变量进行重新提取，公式 10 中 *x* 代表原始数据中 *date* 样本数据。

$$Time = \frac{\text{hour}(x) \times 3600 + \text{minute}(x) \times 60 + \text{second}(x)}{86400} + \text{day}(x) \tag{10}$$

文献[19]对时间变量的处理，如式(11)(12)所示，*NSM* 为 *Date* 样本中时、分、秒转换成总秒数和，*weekstatus* 为星期状态，休息日为 0，工作日为 1。

$$\text{NSM} = \text{hour}(x) \times 3600 + \text{minute}(x) \times 60 + \text{second}(x) \tag{11}$$

$$\text{weekstatus} = \begin{cases} 0 & \text{weekdays}(x) = \text{weekend} \\ 1 & \text{weekdays}(x) = \text{weekday} \end{cases} \tag{12}$$

本文中引入时间戳相对文献[19]中的优点在于，细化了时间粒度，虽然文献[19]中时间戳的引入更容易解释，但是文献[19]中引入的时间特征为 2 个，本文只有一个特征，而且没有时间信息的丢失。在文献[19]中 *NSM* 变量值较大，没有消除量级，影响模型训练的权重，对特征选择时，依据特征重要度排序时有可能会排除掉。本文中引入时间戳，处理后得训练样本集数据如表 2 所示。

表 2　实验数据集

| Date | T | H | Light | CO2 | HR | Time | Occupancy |
|---|---|---|---|---|---|---|---|
| 2015/2/4 18:04 | 23 | 27.125 | 419 | 686 | 0.004714942 | 4.752778 | 1 |
| 2015/2/4 18:06 | 23 | 27.125 | 418.5 | 680.5 | 0.004714942 | 4.752778 | 1 |
| 2015/2/4 18:07 | 23 | 27.2 | 0 | 681.5 | 0.004728078 | 4.754167 | 0 |
| 2015/2/4 18:08 | 22.945 | 27.29 | 0 | 685 | 0.004727951 | 4.754861 | 0 |
| 2015/2/4 18:08 | 22.945 | 27.39 | 0 | 685 | 0.004745408 | 4.755556 | 0 |

(2)基于 *MCMR* 特征选择方法

本文提出基于特征之间冗余度和特征与类别间的相关性相结合的特征选择方法，主要计算过程如下：利用 *Pearson* 系数对样本数据集计算线性相关矩阵 *A*，利用距离相关系数计算特征之间的非相关系数得矩阵 *B*，运用式（3）求得全相关系数矩阵 *C*，相关系数属于[0,1]，其中(0.8,1]属于极强相关，(0.6,0.8]属于强相关，(0.4,0.6]属于中等强相关，(0.2,0.4]属于弱相关，

[0,0.2]属于极弱相关或不相关。可以看出以相关系数度量特征之间的冗余程度，相关系数越大，冗余程度越高。同时基于信息论，计算特征之间互信息率，作为衡量特征与类别之间的相关性，系数越高，特征与类别之间的相关性越高。

通过计算求得全相关矩阵 *C* 如表 3 所示，从表中可以看出 *HumidityRatio* 与 *Humidity* 全相关系数为 0.96，两特征之间具有极强相关性，冗余程度非常高，$CO_2$ 和 *Light* 之间相关系数为 0.596 属于中等相关，其他特征之间相关性较小，冗余程度低。

表 3　全相关矩阵 C

| | T | H | Light | CO2 | HR | Time |
|---|---|---|---|---|---|---|
| T | 1 | 0.335 | 0.558 | 0.518 | 0.334 | 0.442 |
| H | 0.335 | 1 | 0.299 | 0.408 | 0.916 | 0.583 |
| Light | 0.558 | 0.299 | 1 | 0.596 | 0.342 | 0.320 |
| CO2 | 0.518 | 0.408 | 0.596 | 1 | 0.495 | 0.335 |
| HR | 0.334 | 0.916 | 0.342 | 0.504 | 1 | 0.485 |
| Time | 0.442 | 0.583 | 0.320 | 0.335 | 0.485 | 1 |

进一步，为了明确度量各特征含有分类信息大小，进行互信息的计算，然而数值型连续型特征变量无法直接计算互信息[30-31]，所以本文采用 *ChiMerge* 算法进行数值型特征离散化，*ChiMerge* 是最常用的基于卡方的离散化方法，它是一种有监督的、自底向上的数据离散化技术。首先将数据取值范围内的所有值列为一个单独的区间，再递归地找出最佳邻近可合并的区间，通过合并以形成更大的区间。它使用卡方统计量来检测邻近区间相关度，以确定最佳邻近可合并的区间。其中 Time 离散化为 47 类、温度（*T*）离散化为 67 类、湿度（*H*）离散化为 274 类、光照（*Light*）离散化为 56 类、$CO_2$ 浓度（$CO_2$）离散为 239 类、湿度比（*HR*）离散为 718 类，离散化的训练样本数据集如表 4 所示。

表 4　离散化数据集

| T | H | Light | CO2 | HR | Time | Occupancy |
|---|---|---|---|---|---|---|
| 67 | 155 | 21 | 74 | 591 | 1 | 1 |
| 67 | 155 | 22 | 74 | 591 | 1 | 1 |
| 67 | 154 | 21 | 74 | 591 | 1 | 1 |
| 67 | 152 | 21 | 74 | 591 | 1 | 1 |
| 67 | 152 | 21 | 74 | 591 | 1 | 1 |

利用式（3）对离散化的特征,计算特征与类别之间的相关性系数 *w*，得到表 5 所示结果，从表 5 可以看出，特征包含类别信息比率最大的是 *Time* 变量，最小的是 *Temperature* 变量。对相关系数 *w* 进行排序 *Time*>*Light*>*Humidity*>*HumidityRatio*>*CO2*>*Temperature*。

表 5　特征与类别相关矩阵 C

| | T | H | Light | CO2 | HR | Time |
|---|---|---|---|---|---|---|
| Occupancy | 0.245 | 0.303 | 0.460 | 0.289 | 0.293 | 0.552 |

以随机森林算法作为分类器，通过 $\alpha$ 的迭代寻找最有特征子集，初始化设置 $\alpha$=0.4，每次加 0.1,遍历结果如表 6 所示。从表可见变量个数为 5 时，选出的特征子集为最优子集，特征子集为：*Light*，*CO2*，*Humidity*，*Temperature*，*Time*。其中 *HumidityRatio* 特征删除是合理的。

表 6　特征选择结果

| α 数值 | Parameters | Accuracy | |
|---|---|---|---|
| | | accuracy | kappa |
| 0.4 | Light ,Time | 99.27% | 97.81% |
| 0.5 | Light,CO2,H ,Time | 99.38% | 98.15% |
| 0.6 | Light,CO2,H,T , Time | 99.39% | 98.16% |
| | Light,CO2,H,T,HR,Time | 99.37% | 98.11% |

2）实验结果及分析

由数据预处理阶段选择出的最优特征子集，以及样本数据集中时间戳的引入，构建分类模型。获得表 7 的实验结果。

表 7　实验结果比较

| Model | Parameters | Accuracy | | |
|---|---|---|---|---|
| | | training | testing | Testing2 |
| XGBoost_1 | Light,CO2,H,T,HR,Time | 99.41% | 97.67% | 97.65% |
| RF_1 | Light,CO2,H,T,HR,Time | 99.37% | 97.71% | 98.15% |
| 文献 19 | Light,CO2,H,T,HR,NS,WS | 99.36% | 95.53% | 98.06% |
| XGBoost_2 | Light,CO2,H,T,HR | 99.31% | 93.58% | 95.37% |
| RF_2 | Light,CO2,H,T,HR | 99.30% | 94.93% | 97.21% |
| XGBoost_3 | Light,CO2,H,T ,Time | 99.41% | 97.75% | 97.52% |
| RF_3 | Light,CO2,H,T,Time | 99.38% | 97.67% | 97.36% |
| C50 | Light,CO2,H,T,Time | 99.40% | 97.75% | 98.26% |
| SVM | Light,CO2,H,T,Time | 98.71% | 97.90% | 93.64% |
| BP | Light,CO2,H,T,Time | 98.72% | 97.86% | 99.07% |

（1）时间戳对分类精度影响

加入时间戳与没有时间戳的特征组合进行比较，如表 7 所示，XGBoost_1、RF_1 与文献 19 均是没有经过特征选择时加入时间戳，与没经特征选择的，同时也没有引入时间戳的模型相比，XGBoost_1 比 XGBoost_2 在 testing 数据集上分类精度提高了 4.09%， RF_1 比 RF_2 在 testing 数据集上分类精度提高了 2.78%，加入时间戳的模型整体分类精度均得到提高。再者 RF_1 比文献 19 对时间处理，所得分类模型分类精度在每个数据集都高，说明本文的加入时间戳的方法更加合理。

（2）实验结果比较

利用上述步骤获得的最优特征子集，通过调整参数寻得最优分类模型 XGBoost_3、RF_3、C50、SVM、BP 等如表 7 所示，XGBoost_3 在训练样本数据集上的分类精度最高，精度为 99.41%；SVM 在测试样本数据集 1 上的分类精度最高，精度为 97.90%；BP 在测试样本数据集 2 上的分类精度最高，精度为 99.07%。

其中 XGBoost_3 模型的训练集上准确率为 99.41%，在 testing 测试样本的准确率为 97.75%，在 testing2 测试样本集上准确率为 97.52%。RF_3 训练集上的准确率为 99.38%，在 testing 测试样本的准确率为 97.67%，在 testing2 测试样本集上准确率为 97.36%。如图 2 所示，XGBoost_3 模型的分类精度在最优子集上均高于 RF_3 模型的分类精度，在 testing 数据集上差距最大。

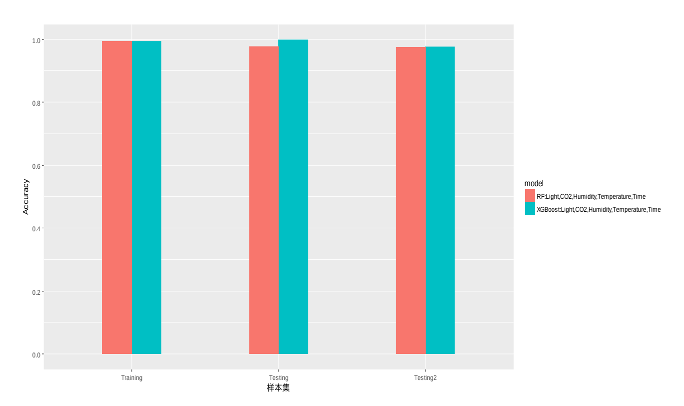



图 2　XGBoost 和 RF 最优特征组合

上述通过分类精度对模型的性能进行了评价，再从时间复杂度对模型的性能进行评价，时间复杂度函数定量地描述一个算法模型的运行时间，对于大量数据的处理，本文在寻求准确的分类精度的同时，仍然追求算法模型处理数据的时间能够相对较短。本文通过 caret 寻优时得到的参值，作为原有算法的参数值，利用原有算法包重新构建分类模型，把 XGBoost 模型和 RF 模型时间复杂度进行对比，通过调用 system.time()函数得到表 8 的时间复杂度，由表可知与传统的随机森林相比，XGBoost 模型用时更少，主要原因在于 XGBoost 算法采用分布式设计，从而降低了时间复杂度。

表 8　模型时间复杂度

| Model | Parameters | running time/s |
|---|---|---|
| XGBoost | Light,CO2,Humidity,Temperature ,Time | 0.4500 |
| RF | Light,CO2,Humidity,Temperature ,Time | 2.2000 |

## 4　结束语

本文主要的工作是对空间占用检测进行研究，在原始数据的基础上，加入时间戳，改变 XGBoost 和 RF 算法无法处理时间变量，实验结果表明，加入时间戳与没有加入时间戳的模型相比分类精度均得到提高，变化最明显的是 XGBoost 在 testing 测试集上分类精度提高了 4.09%，RF 在 testing 数据集上分类精度提高了 2.78%。同时，RF_1 比文献 19 引入时间戳更合理。同时利用 MCMR 特征选择方法进行特征选择，剔除关联度小冗余度高的 HumidityRatio 特征，利用随机森林作为分类器，进行迭代寻优，获取了最优特征子集。通过特征与分类算法构建检测模型得到 XGBoost 在训练样本数据集上的分类精度最高，精度为 99.41%；SVM 在测试样本数据集 1 上的分类精度最高，精度为 97.90%；BP 在测试样本数据集 2 上的分类精度最高，精度为 99.07%。同时 XGBoost 和 RF 分类模型，两者相比，XGBoost 模型分类精度更高，算法时间复杂度更低。

今后的工作将更加深入地研究空间占用检测影响因素，在原有硬传感器的基础上寻找可替代的软传感器获取占用影响变量，减少资源的浪费。同时通过观测得到影响因素数据对空间占用人员个数进行预测和设置资源利用标准对空间进行合理的分配，使得资源得到充足利用。

## 参考文献：

[1] 仇保兴. 我国绿色建筑发展和建筑节能的形势与任务 [J]. 城市发展研究, 2012, 19 (05): 1-7, 11.

[2] 郭萍, 李国刚. 浅析中国当前绿色建筑发展的问题及对策 [J]. 土木建筑与环境工程, 2015, 37 (S1): 96-98.

[3] 徐涛, 王祁. 基于模式识别的传感器故障诊断 [J]. 控制与决策, 2007, (07): 783-786.

[4] 金连文, 钟卓耀, 杨钊, 等. 深度学习在手写汉字识别中的应用综述 [J]. 自动化学报, 2016, 42 (8): 1125-1141.

[5] 杨赛, 赵春霞, 刘凡. 多核学习融合局部和全局特征的人脸识别算法 [J]. 电子学报, 2016, 44 (10): 2344-2350.

[6] Erickson V L, Carreira-Perpi˜nán M Á, Cerpa A E. OBSERVE: occupancy-based system for efficient reduction of HVAC energy [C]// Proc of the 10th International Conference on Information Processing in Sensor Networks. 2011: 258–269.

[7] Erickson V L, M. Á. Carreira-Perpi˜nán, A. E. Cerpa, Occupancy modeling and prediction for building energy management, ACM Trans. Sensor Network, 2014, 10 (3): 42.

[8] Dong B, Andrews B. Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings [C]// Proc of Building Simulation. 2009.

[9] Brooks J, Goyal S, Subramany R, et al An experimental investigation of occupancy-based energy-efficient control of commercial building indoor climate [C]// Proc of the 53rd IEEE Annual Conference on, IEEE, Decision and Control. 2014: . 5680–5685.

[10] Brooks J, Kumar S, Goyal S, et al. Energy-efficient control of under-actuated HVAC zones in commercial buildings [J]. Energy Build, 2015, 93 () 160–168.

[11] Tomastik R, Narayanan S, Banaszuk A, et al. Model-based real-time estimation of building occupancy during emergency egress pedestrian and evacuation dynamics [C]. Berlin: Springer, 2010: 215–224.

[12] Scott J, Brush A B, Krumm J, et al. PreHeat: controlling home heating using occupancy prediction [C]// Proc of the 13th International Conference on Ubiquitous Computing. 2011: 281–290.

[13] Ghai S K, Thanayankizil L V, Seetharam D P, et al. Chakraborty: occupancy detection in commercial buildings using opportunistic context sources [C]// Proc of IEEE Percom Workshops. 2012.

[14] Hailemariam E, R. Goldstein, R. Attar, A. Khan: Real-time occupancy

detection using decision trees with multiple sensor types, in: Proceedings of the 2011Symposium on Simulation for Architecture and Urban Design, Society for Computer Simulation International, San Diego, CA, 2011, pp. 141–148.

[15] Luis M. Candanedo, VÃ©ronique Feldheim. : Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models [J]. Energy and Buildings, 2016, 112 (1): 28-39.

[16] Chen T, He T. Higgs boson discovery with boosted trees [C]// Proc of International Conference on High-Energy Physics and Machine Learning. 2015.

[17] Chen T, Guestrin C. Xgboost: a scalable tree boosting system [J]. arXiv preprint arXiv: 1603. 02754, 2016.

[18] Song R, Chen S, Deng B, et al. eXtreme gradient boosting for identifying individual users across different digital devices [C]// Proc of International Conference on Web-Age Information Management. [S. l. ] : Springer International Publishing, 2016: 43-54.

[19] 姚旭，王晓丹，张玉玺，等．特征选择方法综述 [J]. 控制与决策, 2012, 27 (02): 161-166+192.

[20] 毛勇，周晓波，夏铮，等．特征选择算法研究综述 [J]. 模式识别与人工智能, 2007, 20 (2): 211-218.

[21] 仇利克，郭忠文，刘青，等．基于冗余分析的特征选择算法 [J]. 北京邮电大学学报, 2017, 40 (01): 36-41.

[22] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004: 1205-1224.

[23] 李扬，顾雪平．基于改进最大相关最小冗余判据的暂态稳定评估特征选择 [J]. 中国电机工程学报, 2013, 33 (34): 179-186.

[24] 赵伟卫，李艳颖，赵凤芹，等．基于互信息和随机森林的混合变量选择算法 [J]. 吉林大学学报：理学版, 2017, 55 (04): 933-939.

[25] 徐峻岭，周毓明，等．基于互信息的无监督特征选择 [J]. 计算机研究与发展, 2012, 49 (2): 372-382.

[26] 张振海，李士宁，等．一类基于信息熵的多标签特征选择算法 [J]. 计算机研究与发展, 2013, 50 (6): 1177-1184.

[27] 董红斌，滕旭阳，杨雪．一种基于关联信息熵度量的特征选择方法 [J]. 计算机研究与发展, 2016, 53 (8): 1684-1695.

[28] 薛薇．基于 R 统计分析与数据挖掘 [M]. 北京：中国人民大学出版社, 2014, 1-399.

[29] 周志华．机器学习 [M]. 北京：清华大学出版社, 2016, 1-425.

[30] 赵静娴，倪春鹏，詹原瑞，杜子平．一种高效的连续属性离散化算法 [J]. 系统工程与电子技术, 2009, 31 (01): 195-199.

[31] 杨萍，杨天社，杜小宁，等．一种基于类别属性关联程度最大化离散算法 [J]. 控制与决策, 2011, 26 (04): 592-596.